What is claimed is:

1. A method of detecting duplicate documents, comprising:

receiving a requesting document, such document characterized by a document identifier;

selecting from a plurality of previously received documents a set of documents, if any, sharing the same document identifier, the requesting document and the selected set of documents having associated score information;

updating the selected set of documents with the requesting document in accordance with the score information associated with the requesting document and the selected set of documents; and

determining a representative document for the requesting document and the selected set of documents.

2. The method of claim 1, wherein the document identifier is a fixed length fingerprint of document content of a document characterized by the document identifier.

3. The method of claim 1, wherein the document identifier is a fixed length fingerprint of an address of a document characterized by the document identifier.

4. The method of claim 1, wherein the score information includes a document ranking value indicative of document importance.

5. The method of claim 4, wherein the updating includes adding the requesting document to the selected set of documents if the score information of the requesting document satisfies a predefined insertion condition.

6. The method of claim 5, including evicting another document, if any, from the selected set of documents when a predefined eviction condition is satisfied.

7. The method of claim 6, wherein the evicted document has a lowest document ranking value in the selected set of documents and the total number of documents in the selected set is larger than a predefined value.

8. The method of claim 5, wherein the predefined insertion condition is that the document ranking value of the requesting document is higher than the document ranking value of at least one document in the selected set of documents.

9. The method of claim 1, wherein the determining includes

comparing the score information of the requesting document with that of a particular document from the selected set in accordance with a set of predefined comparison criteria, wherein the particular document was previously determined to be the representative document for the selected set of documents;

selecting the requesting document as the representative document if the set of predefined comparison criteria are met; and

keeping the particular document as the representative document if the set of predefined comparison criteria is not met.

10. The method of claim 9, wherein the set of predefined comparison criteria comprise at least two parameters, one parameter for comparison with an absolute difference of score information between the requesting document and the particular document, and another parameter for comparison with a ratio of score information between the requesting document and the particular document.

11. The method of claim 1, wherein a document is a temporary redirect page comprising a document content, a source document address, and a target document address.

12. A method of detecting duplicate documents in a network crawling system, comprising:

constructing a plurality of tables, each table corresponding to a portion of a document address space, storing information identifying documents having a same document identifier and each identified document having an associated document rank;

receiving a newly crawled document, such document characterized by a document identifier and a document rank;

reading information stored in the plurality of tables to identify a set of documents, if any, sharing the document identifier of the newly crawled document;

updating the information stored in at least one of the tables in accordance with the document ranks of the identified set of documents and the newly crawled document; and

determining a representative document for the newly crawled document and the identified set of documents.

13. The method of claim 12, wherein information identifying the identified set of documents, including a particular document serving as a representative document of the identified set, is stored in one or more tables.

14. The method of claim 12, wherein the determining includes

comparing the document rank of the requesting document with that of the particular document from the identified set in accordance with a set of predefined comparison criteria;

selecting the requesting document as the representative document if the set of predefined comparison criteria are met; and

keeping the particular document as the representative document if the set of predefined comparison criteria is not met.

15. The method of claim 14, wherein the set of predefined comparison criteria comprise at least two parameters, one parameter for comparison with an absolute difference of document ranks between the requesting document and the particular document, and another parameter for comparison with a ratio of document ranks between the requesting document and the particular document.

16. The method of claim 12, wherein the updating includes inserting information identifying the newly crawled document into the at least one table only when a predefined insertion condition is satisfied.

17. The method of claim 16, wherein the predefined insertion condition is that the document rank of the requesting document is higher than the document rank of at least one document in the identified set of documents.

18. A method of detecting duplicate documents in a network crawling system, comprising:

constructing a plurality of tables, each table corresponding to a segment of a document address space, storing information identifying documents having a same document identifier and each identified document having an associated document rank, wherein the plurality of tables comprise N+1 tables where N is an integer greater than one, wherein the

N+1 tables comprise N tables, each generated during a respective phase of a set of N crawling phases, and a current table generated during a current one of the N crawling phases, wherein an oldest one of the N tables was generated during a previous instance of the current crawling phase;

receiving a newly crawled document, such document characterized by a document identifier and a document rank;

reading information stored in the N+1 tables to identify a set of documents, if any, sharing the document identifier of the newly crawled document;

updating the information stored in the current table in accordance with the document rankings of the identified set of documents and the newly crawled document;

determining a representative document for the newly crawled document and the identified set of documents; and

upon completion of the current crawling phase, retiring the oldest one of the N tables.

19. The method of claim 18, wherein the reading comprises reading from a merged table that stores information from a plurality of the N tables, and reading from the current table.

20. The method of claim 18, wherein information identifying the identified set of documents, including a particular document serving as a representative document of the identified set, is stored in one or more tables.

21. A duplicate document detection system, comprising:

one or more central processing units for executing programs;

a network interface for receiving documents; and

a duplicate document detection engine executable by the one or more central processing units, the engine comprising:

a plurality of data structures for storing information of documents, each document characterized by a document identifier and score information, the information stored in the plurality of data structures include the document identifier and score information for each document;

instructions for receiving a requesting document in association with its document identifier and score information;

instructions for selecting from the plurality of data structures a set of documents, if any, sharing the same document identifier as the requesting document;

instructions for generating a new set of documents from the requesting document and the selected set of documents in accordance with their score information; and

instructions for identifying a representative document of the new set of documents.

22. The system of claim 21, wherein the score information for each document includes a document rank metric.

23. The system of claim 21, wherein the plurality of data structures include a data structure for storing information of multiple sets of documents, each set of documents sharing a same document content.

24. The system of claim 21, wherein the plurality of data structures include a data structure for storing information of multiple sets of documents, each set of documents sharing a same document address.

25. The system of claim 21, wherein the document identifier is a fixed length fingerprint of document content of a document characterized by the document identifier.

26. The system of claim 21, wherein the document identifier is a fixed length fingerprint of an address of a document characterized by the document identifier.

27. The system of claim 21, wherein the generating instructions include

sorting the requesting document and the selected set of documents in accordance with a metric included in the score information of the requesting document and selected set of documents; and

selecting a new set of documents, having at most a predefined number of documents, from the requesting document and the selected set of documents based on the sorting result.

28. The system of claim 21, wherein

the score information for each document includes a document rank; and

the identifying instructions include

comparing the document rank of the requesting document with that of a particular document from the selected set of documents in accordance with a set of predefined

comparison criteria, wherein the particular document was previously determined to be the representative document for the selected set of documents;

selecting the requesting document as the representative document for the new set of documents if the set of predefined comparison criteria are met; and

keeping the particular document as the representative document for the new set of documents if the set of predefined comparison criteria is not met.

29. The system of claim 28, wherein the set of predefined comparison criteria comprise at least two parameters, one parameter for comparison with an absolute difference of document rank between the requesting document and the particular document, and another parameter for comparison with a ratio of document rank between the requesting document and the particular document.

30. The system of claim 21, wherein a document is a temporary redirect page comprising a document content, a source document address, and a target document address.

31. A system of detecting duplicate documents in network crawling, comprising:

one or more central processing units for executing programs;

a network interface for receiving documents; and

a duplicate document detection engine executable by the one or more central processing units, the engine comprising:

a plurality of tables, each table corresponding to a portion of a document address space, storing information identifying documents having a same document identifier and each identified document having an associated document rank;

instructions for receiving a newly crawled document, such document characterized by a document identifier and a document rank;

instructions for reading information stored in the plurality of tables to identify a set of documents, if any, sharing the document identifier of the newly crawled document;

instructions for updating the information stored in at least one of the tables in accordance with the document ranks of the identified set of documents and the newly crawled document; and

instructions for determining a representative document for the newly crawled document and the identified set of documents.

32. The system of claim 31, wherein the identified set of documents, including a particular document serving as a representative document of the identified set, are stored in one or more tables.

33. The system of claim 31, wherein the determining includes

comparing the document rank of the requesting document with that of the particular document from the identified set in accordance with a set of predefined comparison criteria;

selecting the requesting document as the representative document if the set of predefined comparison criteria are met; and

keeping the particular document as the representative document if the set of predefined comparison criteria is not met.

34. The system of claim 31, wherein the set of predefined comparison criteria comprise at least two parameters, one parameter for comparison with an absolute difference of document ranks between the requesting document and the particular document, and another parameter for comparison with a ratio of document ranks between the requesting document and the particular document.

35. The system of claim 31, wherein the updating includes inserting information identifying the newly crawled document into the at least one table only when a predefined insertion condition is satisfied.

36. The system of claim 35, wherein the predefined insertion condition is that the document rank of the requesting document is higher than the document rank of at least one document in the identified set of documents.

37. A system for detecting duplicate documents during network crawling, comprising:

one or more central processing units for executing programs;

a network interface for receiving documents; and

a duplicate document detection engine executable by the one or more central processing units, the engine comprising:

a plurality of tables, each table corresponding to a segment of a document address space, storing information identifying documents having a same document identifier and each identified document having an associated document rank, wherein the plurality of tables comprise N+1 tables where N is an integer greater than one, wherein the N+1 tables

comprise N tables, each generated during a respective phase of a set of N crawling phases, and a current table generated during a current one of the N crawling phases, wherein an oldest one of the N tables was generated during a previous instance of the current crawling phase;

instructions for receiving a newly crawled document, such document characterized by a document identifier and a document rank;

instructions for reading information stored in the N+1 tables to identify a set of documents, if any, sharing the document identifier of the newly crawled document;

instructions for updating the information stored in the current table in accordance with the document rankings of the identified set of documents and the newly crawled document;

instructions for determining a representative document for the newly crawled document and the identified set of documents; and

instructions for retiring the oldest one of the N tables upon completion of the current crawling phase.

38. The system of claim 37, wherein the reading comprises reading from a merged table that stores information from a plurality of the N tables, and reading from the current table.

39. The system of claim 37, wherein the identified set of documents, including a particular document serving as a representative document of the identified set, are stored in one or more tables.

40. A computer program product for use in conjunction with a computer system, the computer program product comprising a computer readable storage medium and a computer program mechanism embedded therein, the computer program mechanism comprising:

instructions for constructing a plurality of data structures for storing information of documents, each document characterized by a document identifier and score information, the information stored in the plurality of data structures include the document identifier and score information for each document;

instructions for receiving a requesting document in association with its document identifier and score information;

instructions for selecting from the plurality of data structures a set of documents, if any, sharing the same document identifier as the requesting document;

instructions for generating a new set of documents from the requesting document and the selected set of documents in accordance with their score information; and

instructions for identifying a representative document of the new set of documents.

41. The computer program product of claim 40, wherein the score information for each document includes a document rank metric.

42. The computer program product of claim 40, wherein the plurality of data structures include a data structure for storing information of multiple sets of documents, each set of documents sharing a same document content.

43. The computer program product of claim 40, wherein the plurality of data structures include a data structure for storing information of multiple sets of documents, each set of documents sharing a same document address.

44. The computer program product of claim 40, wherein the document identifier is a fixed length fingerprint of document content of a document characterized by the document identifier.

45. The computer program product of claim 40, wherein the document identifier is a fixed length fingerprint of an address of a document characterized by the document identifier.

46. The computer program product of claim 40, wherein the generating instructions include

sorting the requesting document and the selected set of documents in accordance with a metric included in the score information of the requesting document and selected set of documents; and

selecting a new set of documents, having at most a predefined number of documents, from the requesting document and the selected set of documents based on the sorting result.

47. The computer program product of claim 40, wherein

the score information for each document includes a document rank; and

the identifying instructions include

comparing the document rank of the requesting document with that of a particular document from the selected set of documents in accordance with a set of predefined

comparison criteria, wherein the particular document was previously determined to be the representative document for the selected set of documents;

selecting the requesting document as the representative document for the new set of documents if the set of predefined comparison criteria are met; and

keeping the particular document as the representative document for the new set of documents if the set of predefined comparison criteria is not met.

48. The computer program product of claim 47, wherein the set of predefined comparison criteria comprise at least two parameters, one parameter for comparison with an absolute difference of document rank between the requesting document and the particular document, and another parameter for comparison with a ratio of document rank between the requesting document and the particular document.

49. The computer program product of claim 40, wherein a document is a temporary redirect page comprising a document content, a source document address, and a target document address.

50. A computer program product for use in conjunction with a computer system, the computer program product comprising a computer readable storage medium and a computer program mechanism embedded therein, the computer program mechanism comprising:

instructions for constructing a plurality of tables, each table corresponding to a portion of a document address space, storing information identifying documents having a same document identifier and each identified document having an associated document rank;

instructions for receiving a newly crawled document, such document characterized by a document identifier and a document rank;

instructions for reading information stored in the plurality of tables to identify a set of documents, if any, sharing the document identifier of the newly crawled document;

instructions for updating the information stored in at least one of the tables in accordance with the document ranks of the identified set of documents and the newly crawled document; and

instructions for determining a representative document for the newly crawled document and the identified set of documents.

51. The computer program product of claim 50, wherein information identifying the identified set of documents, including a particular document serving as a representative document of the identified set, is stored in one or more tables.

52. The computer program product of claim 50, wherein the determining includes

comparing the document rank of the requesting document with that of the particular document from the identified set in accordance with a set of predefined comparison criteria;

selecting the requesting document as the representative document if the set of predefined comparison criteria are met; and

keeping the particular document as the representative document if the set of predefined comparison criteria is not met.

53. The computer program product of claim 50, wherein the set of predefined comparison criteria comprise at least two parameters, one parameter for comparison with an absolute difference of document ranks between the requesting document and the particular document, and another parameter for comparison with a ratio of document ranks between the requesting document and the particular document.

54. The computer program product of claim 50, wherein the updating includes inserting information identifying the newly crawled document into the at least one table only when a predefined insertion condition is satisfied.

55. The computer program product of claim 54, wherein the predefined insertion condition is that the document rank of the requesting document is higher than the document rank of at least one document in the identified set of documents.

56. A computer program product of detecting duplicate documents for use in conjunction with a computer system, the computer program product comprising a computer readable storage medium and a computer program mechanism embedded therein, the computer program mechanism comprising:

instructions for constructing a plurality of tables, each table corresponding to a segment of a document address space, storing information identifying documents having a same document identifier and each identified document having an associated document rank, wherein the plurality of tables comprise N+1 tables where N is an integer greater than one, wherein the N+1 tables comprise N tables, each generated during a respective phase of a set

of N crawling phases, and a current table generated during a current one of the N crawling phases, wherein an oldest one of the N tables was generated during a previous instance of the current crawling phase;

instructions for receiving a newly crawled document, such document characterized by a document identifier and a document rank;

instructions for reading information stored in the N+1 tables to identify a set of documents, if any, sharing the document identifier of the newly crawled document;

instructions for updating the information stored in the current table in accordance with the document rankings of the identified set of documents and the newly crawled document;

instructions for determining a representative document for the newly crawled document and the identified set of documents; and

instructions for retiring the oldest one of the N tables upon completion of the current crawling phase.

57. The computer program product of claim 56, wherein the reading comprises reading from a merged table that stores information from a plurality of the N tables, and reading from the current table.

58. The computer program product of claim 56, wherein the identified set of documents, including a particular document serving as a representative document of the identified set, are stored in one or more tables.